Seminar 7, Some Derivations of Mean-Field Updates

Evgenii Egorov, egorov.evgenyy@ya.ru

September 15, 2020

Goals of the note:

1. Show how to derive updates on simple models example
2. Provide general derivation for conditional-conjugate model

1 Some derivations for simple models

1.1 Normal-Gamma Model

We consider the following model, for $x_n \in \mathbb{R}$, $X = \{x_n\}_{n=1}^N$, $\theta = (\mu, \lambda)$ joint density:

$$p(X, \mu, \lambda) = \left[\prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1})\right] \mathcal{N}(\mu; m, (\beta\lambda)^{-1}) G(\lambda; a_0, b_0).$$

It will be useful to write the log of the joint density:

$$\log p(X, \mu, \lambda) = \left[\sum_{n=1}^{N} \frac{1}{2}\log\lambda - \frac{\lambda}{2}(\mu - x_n)^2\right] + \frac{1}{2}\log(\beta\lambda) - \frac{\beta\lambda}{2}(\mu - m)^2 + (a_0 - 1)\log\lambda - b_0\lambda.$$

We consider the following approximation of the posterior: $p(\mu, \lambda|X) = q(\lambda)\, q(\mu)$. The general mean-field update equation: $\log q(\theta_j) \propto^+ \langle \log p(X, \theta)\rangle_{q(\theta_{-j})}$. We derive iterative updates for $q(\mu)$ and $q(\lambda)$ using it, starting with $q(\mu)$:

$$\log q(\mu) \propto^+ \langle\lambda\rangle_{q(\lambda)} \left[-\frac{1}{2}\sum_{n=1}^{N}(\mu - x_n)^2\right] + \langle\lambda\rangle_{q(\lambda)} \left[-\frac{\beta}{2}(\mu - m)^2\right].$$

From here we can "recognize" normal distribution, as we have log-density as sum of the quadratic forms. Hence, the distribution defined by it two first moments, which we find with help of the mode and hessian.

$$\nabla_\mu \log q(\mu) = 0 \Longleftrightarrow \sum_{n=1}^{N} (\mu - x_n) + \beta(\mu - m_0) = 0,$$

$$\boxed{m' = \frac{1}{N+\beta}\left[\sum_{n=1}^{N} x_n + \beta\mu\right]}, \boxed{\lambda'^{-1} = -[\nabla_\mu^2 \log q(\mu)]^{-1} = [(N+\beta)\langle\lambda\rangle_{q(\lambda)}]^{-1}}.$$

Hence, $\boxed{q(\mu) = \mathcal{N}(\mu|m', \lambda'^{-1})}$. And for $q(\lambda)$:

$$\log q(\lambda) \propto^{+} \frac{N}{2}\log\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}\langle(\mu - x_n)^2\rangle_{q(\mu)} + \frac{1}{2}\log(\beta\lambda) - \frac{\beta\lambda}{2}\langle(\mu - m)^2\rangle + (a_0 - 1)\log\lambda - b_0\lambda =$$

$$= \left(\frac{N}{2} + \frac{1}{2} + a_0 - 1\right)\log\lambda - \lambda\left(\frac{1}{2}\sum_{n=1}^{N}\langle(\mu - x_n)^2\rangle_{q(\mu)} + \frac{1}{2}\beta\langle(\mu - m_0)^2\rangle_{q(\mu)} + b_0\right).$$

Again, we can "recognize" gamma distribution, thus $\boxed{q(\lambda) = \text{Gamma}(a_0', b_0')}$:

$$a_0' = \boxed{a_0 + \frac{N}{2} + \frac{1}{2}},$$

$$b_0' = \boxed{b_0 + \left(\frac{1}{2}\sum_{n=1}^{N}\langle(\mu - x_n)^2\rangle_{q(\mu)} + \frac{1}{2}\beta\langle(\mu - m)^2\rangle_{q(\mu)}\right)}.$$

For the Normal-Gamma model, we also can derive the exact posterior $p(\mu, \lambda|X) = p(\mu|\lambda, X)p(\lambda|X)$. Let's do this in order to compare with mean-field approximation, derived above.

$$\int p(X|\mu,\lambda)p(\mu|\lambda)p(\lambda)\ d\mu = p(\lambda)\left[\prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\lambda^{-1})\right]\mathcal{N}(\mu; m, (\beta\lambda)^{-1})\ d\mu. \text{ Hence:}$$

$$p(\mu|\lambda, X) = \frac{p(X|\mu,\lambda)p(\mu|\lambda)p(\lambda)}{p(\lambda)\int p(X|\mu,\lambda)p(\mu|\lambda)\ d\mu} = \frac{1}{Z}\left[\prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\lambda^{-1})\right]\mathcal{N}(\mu; m, (\beta\lambda)^{-1}),$$

which leads to the normal distribution for the $p(\mu|\lambda, X)$. Similarly to the derivations for the mean-field, we obtain:

$$\boxed{p(\mu|\lambda, X) = \mathcal{N}\left(\mu|\tfrac{1}{N+\beta}\left[\sum_{n=1}^{N} x_n + m\right], [\lambda(N+\beta)]^{-1}\right)},$$

$$\boxed{p(\lambda|X) = G(\lambda; a_0 + \tfrac{N}{2}, b')}.$$

1.2 Bayesian GMM

Consider the following model:

$$p(X,z,\pi,\mu,\Lambda) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\left(\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})\pi_k\right)^{z_{nk}}\right] \mathrm{Dir}(\pi|\alpha_0)\prod_{k=1}^{K}\mathcal{N}(\mu_k|m_0,(\beta\Lambda_k)^{-1})W(\Lambda_k|W_0,\mu_0).$$

We consider following approximation:

$$p(z,\pi,\mu,\Lambda|X) = q(z)q(\pi,\mu,\Lambda).$$

Let's take a look on logarithm of the joint density:

$$\log p(X,z,\pi,\mu,\Lambda) = \sum_{n,k} z_{nk}\left(\frac{1}{2}\log|2\pi\Lambda_k| - \frac{1}{2}(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k) + \log\pi_k\right) + \sum_k(\alpha_0-1)\log\pi_k +$$
$$+\sum_k \frac{1}{2}\log|2\pi\beta\Lambda_k| - \frac{\beta}{2}(\mu_k-m_0)^T\Lambda_k(\mu_k-m_0) + \frac{\mu_0-d-1}{2}\log|\Lambda_k| - \frac{1}{2}\mathrm{tr}(\lambda_k W_0^{-1}).$$

From it, we can see (that of cause not surprise) that we have:

$$q(z)q(\pi,\mu,\Lambda) = q(z)q(\pi)\prod_{k=1}^{K}q(\mu_k,\Lambda_k).$$

Let's denote $q(z_{nk}=1) = r_{nk}$ and focus on $q(\pi)$.

$$\log q(\pi) \propto \langle \log p(x,z,\pi,\mu,\Lambda)\rangle_{q(z)q(\mu,\Lambda)} = \sum_k(\alpha_0 - 1 + \sum_n r_{nk})\log\pi_k.$$

Hence, we have Dirichlet:

$$q(\pi) = \mathrm{Dirichlet}(\pi|\alpha_0 + \sum_n r_{nk}).$$

And

$$\mathbb{E}\pi_k = \frac{\alpha_k}{\sum_k \alpha_K} = \frac{\alpha_0 + \sum\limits_n r_{nk}}{K\alpha_0 + N}.$$

Hence, for small $\sum\limits_n r_{nk}$, we obtain probability of assignment around 0. So, you can take $K = N$ and obtain a small number of the clusters. Note, that similarly with RVM implementation, you can reduce K during training iterations.

2 More general derivations for Exp. Families

For mean-field updates, we use following updates:

$$\begin{aligned} p(x_{1:n}, z_{1:m}) &= p(x_{1:n}) \prod_{j=1}^{m} p(z_j | z_{1:j-1}, x_{1:n}), \\ \log q(z_j) &\propto^{+} \mathbb{E}_{-j} \log p(z_j | z_{-j}, x_{1:n}). \end{aligned}$$

So, if we could define $p(z_j|z_{-j}, x_{1:n})$ to get posterior approximation. We consider the exponential family for this conditionals:

$$\begin{aligned} p(z_j|z_{-j}, x_{1:n}) &= \exp\left(\langle \lambda(z_{-j}, x_{1:n}), \phi(z_j) \rangle - F(\lambda(z_{-j}, x)\right), \\ \log p(z_j|z_{-j}, x_{1:n}) &= \langle \lambda(z_{-j}, x_{1:n}), \phi(z_j) \rangle - F(\lambda(z_{-j}, x_{1:n}), \\ \log q(z_j) &\propto^{+} \mathbb{E}_{-j} \log p(z_j|z_{-j}, x_{1:n}) = \langle \mathbb{E}_{-j}\lambda(z_{-j}, x_{1:n}), \phi(z_j) \rangle, \\ q(z_j) &= \exp\left(\langle \mathbb{E}_{-j}\lambda(z_{-j}, x_{1:n}), \phi(z_j) \rangle - F(\mathbb{E}_{-j}\lambda(z_{-j}, x_{1:n}))\right). \end{aligned}$$

Hence, we have $q(z_j)$ as the same distribution as $p(z_j|z_{-j}, x_{1:n})$, but instead $\lambda(z_{-j}, x_{1:n})$ natural parameter is equal $\mathbb{E}_{-j}\lambda(z_{-j}, x_{1:n})$.